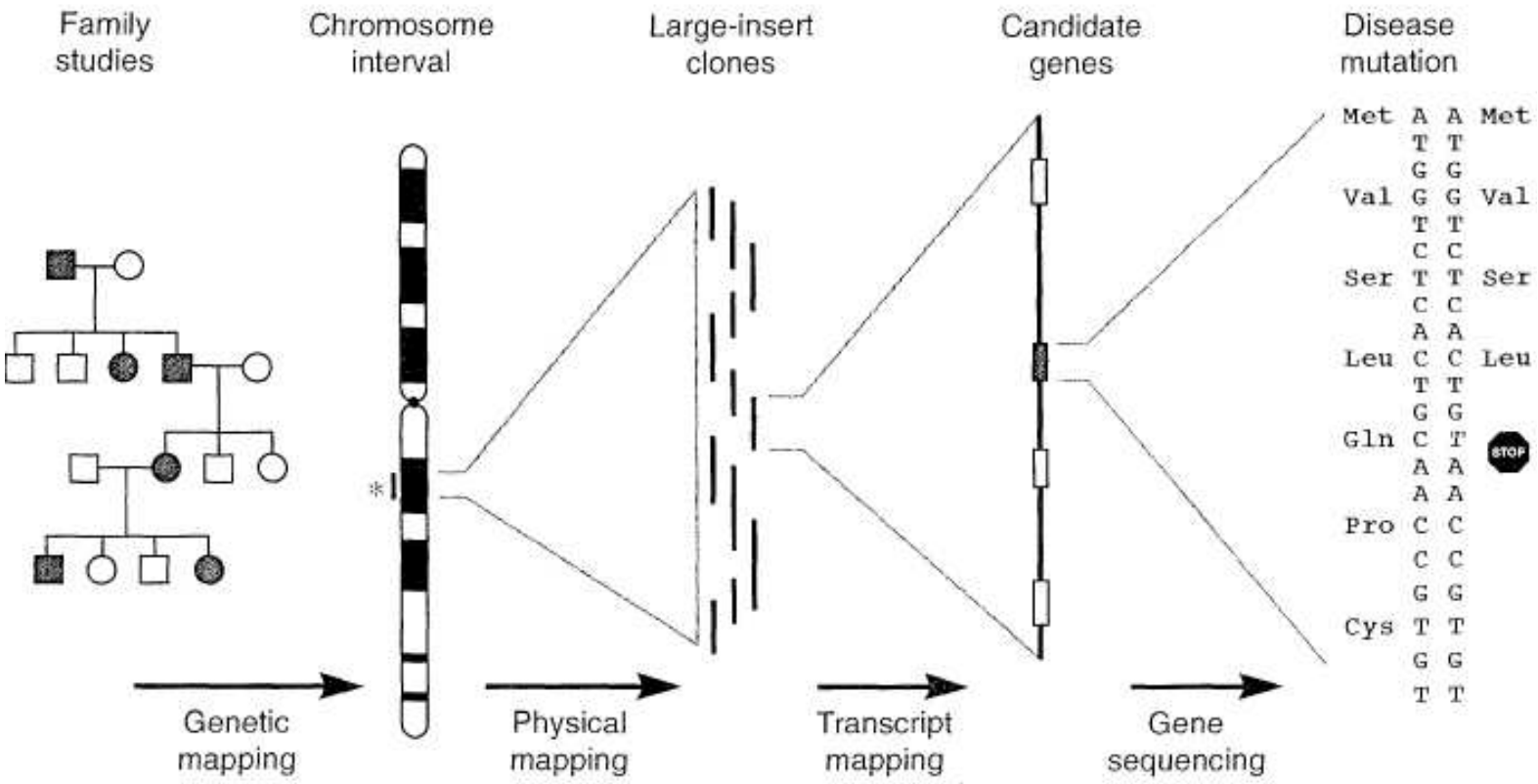
IV. Association analysis I

- Scope and concepts
- Population-based association analysis
- Family-based association analysis
- Case studies
- Practice: EHPLUS, ETDT, PLINK, HaploView, R, SAS, Stata

Scope

- **Genetic association study** is the search for genetic markers that occur at a different frequency between cases and controls, or vary with a quantitative trait.
- However, allele frequencies could vary between cases and controls for many reasons,
  - The allele itself is a cause of the disease
  - The allele is in linkage disequilibrium (LD) with a disease causing gene
  - This is due to an artefact of population stratification
- A simple test of frequency difference between unrelated cases and controls can be done with Pearson chi-squared statistic but more generally, this could also be done with family data.

Steps in positional cloning



Positioning of disease loci is localized through linkage studies, to be followed by physical mapping and sequencing (Schuler et al. 1996)

Change of paradigms

- Traditional (e.g. segregation and linkage) methods in human gene-trait study gradually move towards association study. They do not solely rely on family data.
- Large-scale studies as with good coverage of the genome are required to maintain statistical power, so it is costly but hugely important. However, it is notable with SNPs,
  - The mean density SNPs is approximately one per kilobase in the human genome, so that a SNP library will be sufficient for various study designs.
  - The low mutation rate per generation of SNPs makes them appropriate for association studies.
  - Findings regarding Mendelian disorders suggest a subset of SNPs are functionally important for complex traits.
  - Low-cost, high-throughput, automatic genotyping methods are available.

Concepts

- Coalescence
- Linkage disequilibrium
- Population-based and family-based designs
- Population stratification
- Gene-gene and gene-environment interactions

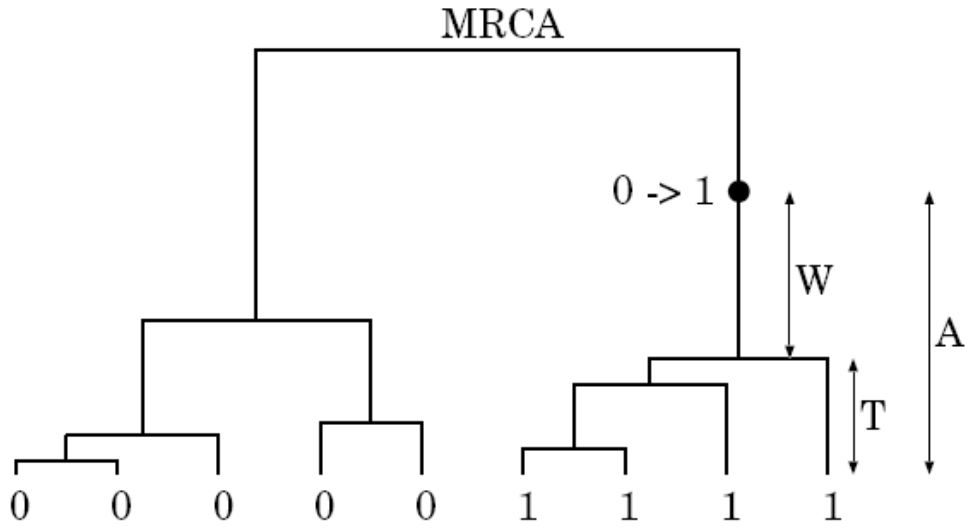
Coalescent models

- It is of interest for many reasons (Fu & Li 1999)
  - It is a sample-based theory, proposing that description of a sample is more relevant that that of a whole population.
  - It is efficient and motivates many algorithms for simulating population samples under various population genetics models.
  - It is particularly suitable for molecular data, such as DNA sequence sample.
- Consider a sample of n sequences of DNA from a population and no recombination. They are connected by a single phylogenetic tree or genealogy, the root of the tree is the most recent common ancestor (MRCA). When n is small relative to total population size N, to a first approximation, only two lineages coalesce at each event and the distribution of times between successive coalescent events has a geometric distribution

$$[1-i(i-1)/4N]^{t-1}[i(i-1)/4N] \approx \exp[-i(i-1)t/4N][i(i-1)/4N]$$

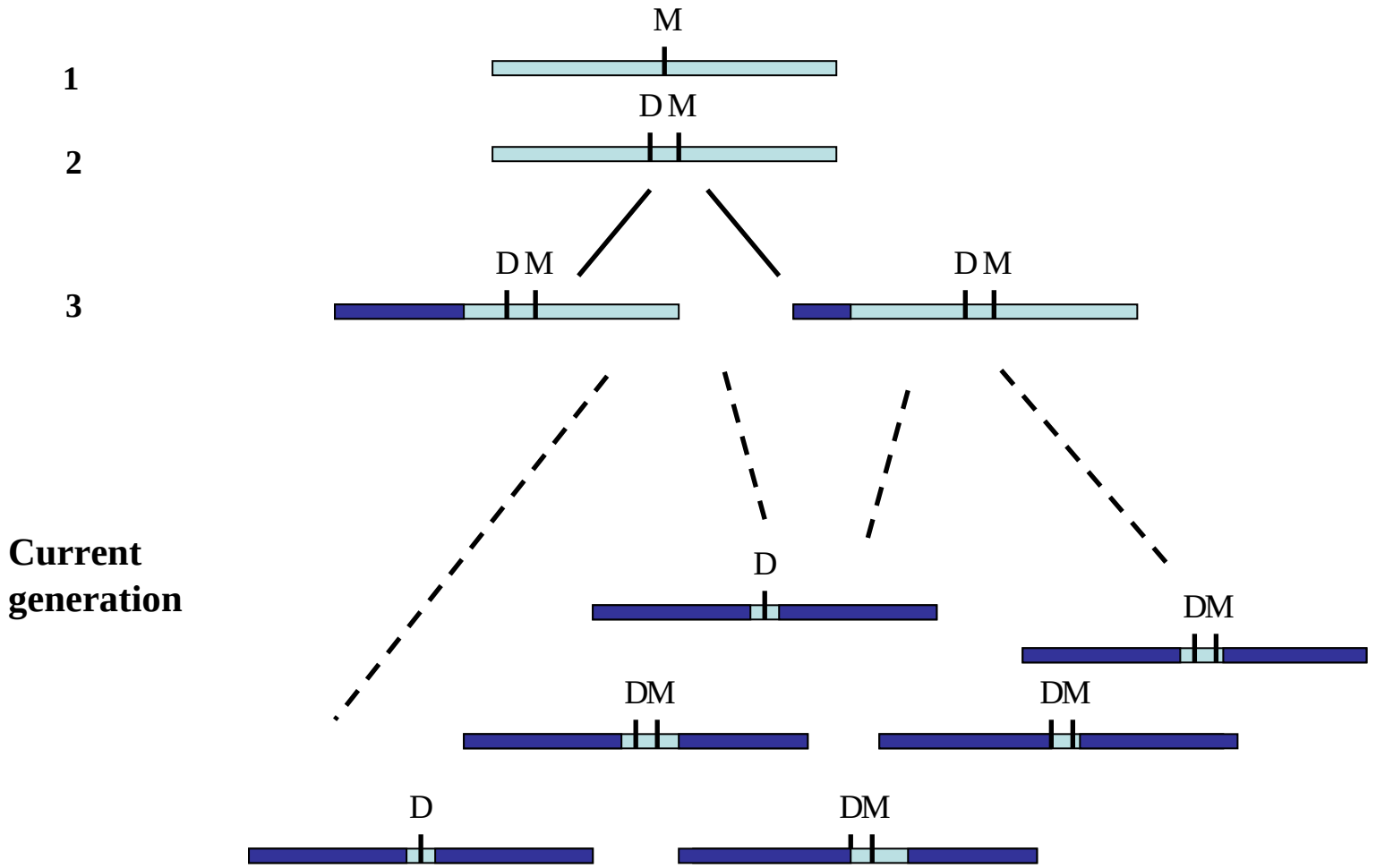
- This implies that for i alleles, the probability of no coalescence for the first t-1 generations followed by coalescence at t^{th} generation.

MRCA



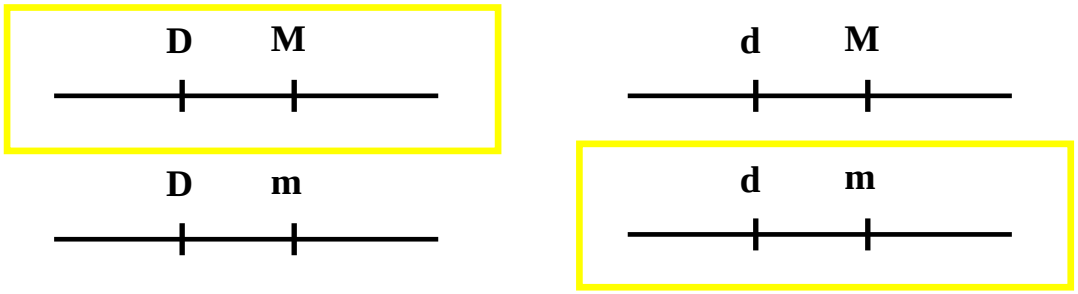
Allelic genealogy where a single mutation (solid circle) $0 \longrightarrow 1$ gives rise to a sample of 4 chromosomes carrying the new 1 allele. Indicated are MRCA of the whole sample at time T_{MRCA}, the MRCA of all chromosomes with the 1 allele at time T, the age of the mutation A and the difference between the allelic MRCA and the allelic age W.

Linkage disequilibrium



Testing for association with SNPs

Haplotypes



- Power to detect association with marker M depends on
  - frequency of D, P(D)
  - frequency of M, P(M)
  - LD between D and M, P(M|D)
  - risk of disease for genotypes Dd, DD, γ, γ2

Measures of LD

- The expectation of $D = 0$

$$D_{AB} = f_{AB} - f_A f_B$$

- r^2 correlation coefficient
  - Range [0,1]
  - Hill & Robertson (1968)

$$r^2_{AB} = \frac{D^2_{AB}}{f_A f_a f_B f_a} = \rho^2_{AB}$$

- D’
  - Range [0,1]
  - Lewontin (1964)

$$|D'_{AB}| = \text{if } (D > 0)\ \frac{D_{AB}}{\min(f_A f_b, f_a f_B)}$$

$$\text{else } \frac{-D_{AB}}{\min(f_A f_B, f_a f_b)}$$

- Odds-ratio formulation
  - Devlin & Risch (1995)

$$\delta_{AB} = \frac{D_{AB}}{f_B f_{ab}}, D_{AB} > 0$$

- For the Malecot model

$$E(D') = Ae^{-td} + B, (1-\theta) \approx e^{-d}$$

Variances of r and D’ available from 2LD and LD22 in R/gap

Population-based association tests

- Genotypewise versus allelewise analysis
- Armitage trend test
- Relationship with logistic regression
- Bayesian statistics
- Linear regression

Cochran-Armitage Trend Tests

- Assuming the sample to be typed at a SNP marker of interest, we can represent genotype data in a 2 x 3 contingency table.
- The Cochran-Armitage trend test of association between disease and the marker SNP is given by

$$X^2 = \frac{\left[\left(p_{2A} + \frac{1}{2}p_{1A}\right) - \left(p_{2U} + \frac{1}{2}p_{1U}\right)\right]^2}{\left(\frac{1}{n_{\cdot A}} + \frac{1}{n_{\cdot U}}\right)\left(\frac{1}{n_{\cdot\cdot}^2}\right)\left[n_{\cdot\cdot}\left(\frac{1}{4}n_{1\cdot} + n_{2\cdot}\right) - \left(\frac{1}{2}n_{1\cdot} + n_{2\cdot}\right)^2\right]}$$

where

$$p_{ij} = \frac{n_{ij}}{n_{\cdot j}}$$

- X^2 has χ^2 distribution with 1 degree of freedom under null hypothesis.

	Cases	Controls	Total
MM	n_{2A}	n_{2U}	$n_{2\cdot}$
Mm	n_{1A}	n_{1U}	$n_{1\cdot}$
mm	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

- Odds ratio for allele M relative to allele m

$$\psi_{M|m} = \frac{\left(\frac{n_{1A}n_{0U}}{n_{0\cdot}+n_{1\cdot}}\right) + \left(\frac{n_{2A}n_{1U}}{n_{1\cdot}+n_{2\cdot}}\right) + \left(\frac{4n_{2A}n_{0U}}{n_{0\cdot}+n_{2\cdot}}\right)}{\left(\frac{n_{0A}n_{1U}}{n_{0\cdot}+n_{1\cdot}}\right) + \left(\frac{n_{1A}n_{2U}}{n_{1\cdot}+n_{2\cdot}}\right) + \left(\frac{4\left[n_{2A}n_{2U}n_{0A}n_{0U}\right]^{1/2}}{n_{0\cdot}+n_{2\cdot}}\right)}$$

- Affected individual $\psi_{M|m}^2$ times more likely to have marker genotype MM than mm, and $\psi_{M|m}$ times more likely to have genotype Mm than mm.

Allele-based Single-locus Tests

- Each individual now contributes ***two*** counts to the contingency table, one for each allele in their marker genotype.
- Assuming the sample to be typed at a marker SNP of interest, we can represent genotype data in a 2 x 2 contingency table.
- To test the null hypothesis of no disease-marker association

- Where $X^2 = \sum_{i=0,1} \sum_{j=A,U} \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]}$

$$E[n_{ij}] = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

- X^2 has χ^2 distribution with 1 degree of freedom under null hypothesis.

	Cases	Controls	Total
M	n_{1A}	n_{1U}	$n_{1\cdot}$
m	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

- Odds ratio for allele M relative to m

$$\psi_{M|m} = n_{1A} n_{0U} / n_{0A} n_{1U}$$

- Allele M is $\psi_{M|m}$ times more likely to be carried by an affected individual than allele m.
- Assumes multiplicative disease risks and Hardy-Weinberg equilibrium at SNP in cases and controls.

Logistic regression

- For the 2x3 table

Z	0	1	2
Cases	s_0	s_1	s_2
Controls	r_0	r_1	r_2

- The logistic regression model is

$$L(\theta) = P(Y|Z,\mu,\gamma) = \prod_{i=1}^{N} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

where

$$\theta = (\mu,\gamma) \qquad \log\frac{p_i}{1-p_i} = \mu + \gamma Z_i \quad p_i = \frac{e^{\mu+\gamma Z_i}}{1+e^{\mu+\gamma Z_i}}$$

The score test statistic (Marchini et al. 2007)

- A test for γ=0 can be done with score test statistic

$$S = U_\gamma^T I^{-1} U_\gamma = \frac{N(N_2(s_1 + 2s_2) - N_1(r_1 + 2r_2))}{N_1 N_2(s_1 + r_1 + 4(s_2 + r_2) - (s_1 + r_1 + 2(s_2 + r_2))^2)}$$

  which is the Cochran-Armitage trend test.

- A general 3-parameter model

$$\log \frac{p_i}{1 - p_i} = \mu + \gamma I(Z_i = 1) + \phi I(Z_i = 2)$$

- has score test statistic for (γ=φ=0) is $S = U^T I^{-1} U$ where

$$U = \left(s_1 - \frac{N_1}{N}(s_1 + r_1), s_2 - \frac{N_1}{N}(s_2 + r_2)\right)^T$$

$$I = \frac{N_1 N_2}{N^3} \begin{pmatrix} N(s_1 + r_1) - (s_1 + r_1)^2 & -N(s_1 + r_1)(s_2 + r_2) \\ -N(s_1 + r_1)(s_2 + r_2) & N(s_2 + r_2) - (s_2 + r_2)^2 \end{pmatrix}$$

Continuous trait

- For normally distributed trait, a linear regression can be conducted. It is worthwhile to note that under this model, the proportion of variance explained (R^2) is also the square of the product-moment (Pearson) correlation coefficient.
- More generally, the trait-marker association can be cast in the generalized (latent) linear (mixed) model framework. Note also that for case-control data, this is also a prospective model, it might be more appropriate to use retrospective model.
  - Prospective model: P(Y|G)
  - Retrospective model: P(G|Y)

Unknown phase in multilocus analysis

- Haplotype is a collection of alleles from neighbouring loci.
- There are $2^{(L-1)}$ possible phases if L loci are heterozygous
- Let $\theta=(\theta_1, \theta_2, \ldots, \theta_J)$ be the vector of haplotype frequencies. Assuming Hardy-Weinberg equilibrium,

$$P(g) = P(h) = \prod_{j-1}^{J} \binom{2}{h_j} \theta^{h_j} = \binom{2}{h_j} \prod_{j-1}^{J-1} \binom{2}{h_j} \varphi^{h_j} \left(1 + \sum_{j=1}^{J-1} \varphi_j\right)^{-1}$$

and

$$\varphi_j = \theta_j / 1 - \sum_{j=1}^{J-1} \theta_j$$

The combined model

The complete data log-likelihood for the *i*th individual ignoring some constants is given by LMM and more general GLMM

$$l_i = \frac{y_i \eta - b(\eta)}{a(\phi)} + \sum_{j=1}^{J-1} h_{ij} \log \varphi_j - 2\log(1 + \sum_{j=1}^{J-1} \varphi_j)$$

where $\eta = X\beta + Z\gamma$. If t_i indicates both genetic and environmental effects, then an EM algorithm involves,

$$P(h_{ij} h_{ij'} \mid d_i^{(o)}) = \frac{f(y_i \mid t_i, z_i) p(g_i)}{\sum_{g \in S(G)} f(y_i \mid t_i, z_i) p(g_i)}$$

Lake et al. (2003) Hum Hered

Gene counting

The original data

ID	label	mar1	mar2
1	1	1 2	1 2
2	0	1 2	1 2
3	0	1 2	1 1
4	1	1 1	1 2
5	0	1 1	1 2
6	0	1 2	1 1
7	1	2 2	2 2
8	0	1 2	2 2
9	0	2 2	1 2
10	1	1 1	1 2
11	0	1 1	1 2
12	0	1 2	1 1

(a)

Data file required by **EH**

EH

```
case.dat
 2   2
  0   2   0
  0   1   0
  0   0   1

control.dat
 2   2
  0   2   0
  3   1   1
  0   1   0
```

(b)

modified

```
2  2
 2  2  2
 4  0  3
 5  1  1
 6  0  1
 8  0  1
 9  1  0
```

columns after row 1 in (c):
1 genotype identifier
2 count of cases
3 count of controls

(c)

Data files required by **EH** before and after revision. **(a)** The raw dataset contains subject id, case-control label(1=case, 0=control), two biallelic markers mar1 and mar2. **(b)** Data files for **EH** (case.dat and control.dat) and **(c)** The data file for the modified program. Note lines with identifiers 1, 3, 7 are omitted.

Two-locus (SNP) genotype-haplotype relationship

SNP1 \ SNP2	1/1	1/2	2/2	missing
1/1	n_0	n_1	n_2	m_1'
1/2	n_3	n_4	n_5	m_2'
2/2	n_6	n_7	n_8	m_3'
missing	m_1	m_2	m_3	

SNP1 \ SNP2	1	2
1	h_{11}	h_{12}
2	h_{21}	h_{22}

Genotypes as a function of haplotypes

Marker 1	Marker 2 1/1	1/2	2/2	t'
1/1	h_{11}^2	$2h_{11}h_{12}$	h_{12}^2	t'_1
1/2	$2h_{21}h_{11}$	$2h_{21}h_{12} + 2h_{22}h_{11}$	$2h_{22}h_{12}$	t'_2
2/2	h_{21}^2	$2h_{21}h_{22}$	h_{22}^2	t'_3
t	t_1	t_2	t_3	

Haplotype association

- Log-likelihood = $\sum n \ln(p)$
  - where *n,p* are the genotype count and probability
  - H_0:*p* is made of independent haplotype frequencies;
  - H_1:*p* is formed by haplotype frequencies

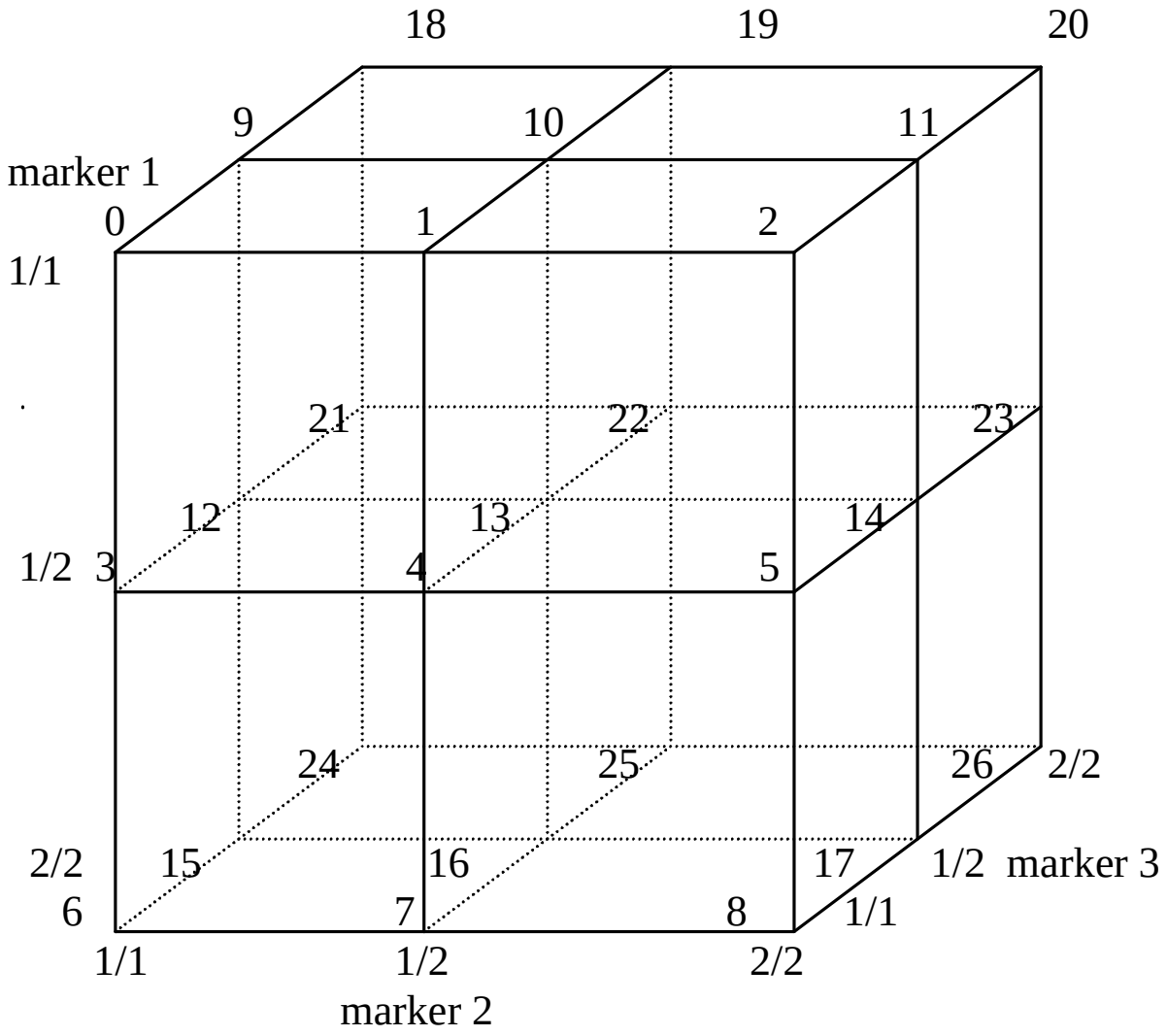
  Therefore, association from a contingency table can be partitioned into general, haplotype and residual components (ASSOCIATE, 2LD).

- In the case of missing data, the likelihood function has the following form,

$$l = \sum_{i=1}^{9} n_i \ln(g_i) + \sum_{j=1}^{3} m_j \ln(t_j) + \sum_{k=1}^{3} m'_k \ln(t'_k)$$

- Log-likelihood ratio test provides a test of genetic association.

A three-locus model



Bayesian vs classical statistics

Models $P(Data|\theta_1, M_1)$ $P(Data|\theta_2, M_2)$

Priors $P(\theta_1|M_1)$ $P(\theta_2|M_2)$

$$\text{Bayes Factor} = \frac{P(Data|M_1)}{P(Data|M_2)} = \frac{\int P(Data|\theta_1, M_1)P(\theta_1|M_1)d\theta}{\int P(Data|\theta_2, M_2)P(\theta_2|M_2)d\theta}$$

$P(Data|M_1)$ and $P(Data|M_2)$ are marginal likelihoods.

$$\text{Likelihood Ratio} = \frac{\max_{\theta_1} P(Data|\theta_1, M_1)}{\max_{\theta_2} P(Data|\theta_2, M_2)}$$

BF is interpreted as the factor by which the prior odds of association are changed in light of the data to produce the posterior odds of association: Posterior odds=BF x Prior odds. For instance M_1 denotes the model in which the SNP is associated with an additive effect on the log-odds scale, in contrast to the null model M_0 of no association, then $\theta_1=(\mu,\gamma)$, $\log(p_i/(1-p_i)= \mu+ \gamma Z_i$, and $\theta_0=(\mu,\gamma)$, $\log(p_i/(1-p_i)= \mu$. We can take $\mu \sim N(0,1)$, with N(.,.) for γ.

Marchini et al. (2007) Nat Genet

Bayes factor

- Given the prior effect size as $N(0,\sigma^2)$
- $BF = (1+c^2)^{-0.5}\exp(z^2/2(1+c^2))$ with $c = \sigma/s$
  - P depends on z
  - BF depends on s, therefore sample size
  - We expect most null hypotheses to be true, i.e., the prior odds are against association, therefore a large BF is required and a small p-value, a compromise has been suggested to be when $N \approx 1/\sigma^2$, namely N is proportional to $1/(\text{effect size})^2$
- In a genomewide association study the prior odds against association for any gene might be approximately 3,000:1, requiring a BF of at least 10^4, this corresponds to p–values $<10^{-6}$

Population stratification

- It has been one of the major concerns for genetic association studies, especially with population-based sample.
- A set of statistical methods has been proposed, including
  - Genomic control
  - Structured association
  - Family data
- Genomic control uses frequentist or Bayesian approaches to adjust for evidence of association by introducing an inflation factor for the statistics produced by association testing. It is simple to use.
- With availability of large number of SNPs, it is possible to infer hidden population structure through standard multivariate analyses such as principal component and correspondence analyses.

Family-based association tests

- It has been used as alternatives to tests for population-based samples and appropriate for family data.
- It involves the so-called transmission disequilibrium test (TDT)
  - discrete trait
  - quantitative trait

TDT for discrete trait

- For a large sample of trios (with varying genotypes) we can form a table. H0: transmissions of allele 1 from a 1/2 parent are equally likely as transmissions of allele 2, (b = c). The test of association is McNemar's: $(b - c)^2/[b + c]$ with chi-squared distribution, (b + c) is an estimate of the variance of $(b-c)^{2;}$ so that data from only 2 cells (b, c) are used.
- Transmissions from homozygous parents and (1/1 or 2/2) are ignored in the analysis.
- Alternative, more powerful statistic such as HHRR, but is not protected against population stratification
- It has been extended to multiallelic markers, quantitative traits, pedigree data and only siblings.

	U		
T		1	2
	1	a	b
	2	c	d

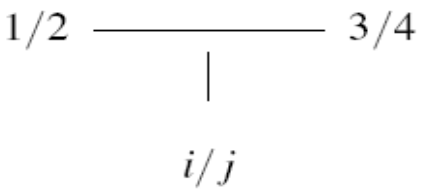
T=transmitted allele

U=untransmitted allele

The relative risk estimate (RR) b/c with variance estimate Var(log(RR))=1/b+1/c

Conditioning on parental genotype

- Consider parents with genotypes $1/2$ and $3/4$. An unselected offspring can have one of four genotypes with equal probability:



$$i/j = \begin{matrix} 1/3 \\ 1/4 \\ 2/3 \\ 2/4 \end{matrix}$$ each with probability 0.25

- Denoting the risk of disease in a subject with genotype i/j by $\pi\theta_{i/j}$, where θ denotes a *genotype relative risk* then, if we choose the family with an *affected* offspring, then the probability that the offspring has genotype i/j is

$$\frac{\theta_{i/j}}{\theta_{1/3}+\theta_{1/4}+\theta_{2/3}+\theta_{2/4}}$$

- Writing $\theta_{i/j}=\phi_i\phi_j$, the conditional probability becomes

$$\frac{\phi_i\phi_j}{\phi_1\phi_3+\phi_1\phi_4+\phi_2\phi_3+\phi_2\phi_4}=\frac{\phi_i}{\phi_1+\phi_2}\times\frac{\phi_j}{\phi_3+\phi_4}$$

TDT for quantitative trait (Wang and Cohen 1999)

- A multiallelic TDT statistic is defined as

$$TDT_m = \frac{m-1}{m} \sum_{i=1}^{m} \frac{(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot i})^2}{\left(\frac{1}{n_{i\cdot}} + \frac{1}{n_{\cdot i}}\right) S_i^2}$$

- where

$$\bar{Y}_{i\cdot} = \frac{1}{n_{i\cdot}} \sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{k=1}^{n_{ij}} Y_{ijk}, \quad \bar{Y}_{\cdot i} = \frac{1}{n_{\cdot i}} \sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{k=1}^{n_{ji}} Y_{jik},$$

$$S_i^2 = \frac{1}{n_{i\cdot} + n_{\cdot i} - 2} \left[\sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{i\cdot})^2 + \sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{k=1}^{n_{ji}} (Y_{jik} - \bar{Y}_{\cdot i})^2 \right]$$

Epidemiological criteria for gene-association

(Tabor et al. 2002)

- Biological plausibility of association and its consistency with existing knowledge about biology and disease aetiology are evaluated. Is the candidate gene likely to be involved in the phenotype? Are the single-nucleotide polymorphisms (SNPs) likely to have functional effects on the protein?
- The strength of the association between the risk factor and the disease is examined. When considering multiple SNPs in a candidate gene, the ones with strongest association are most likely to be causally related.
- The dose–response relationship of the association is considered. For example, individuals with two copies of a variant might be at greater risk of disease than individuals with one copy of the variant.
- The consistency of the association across past and future studies, and across different populations, is an important consideration. Consistent replication in different populations is strong evidence of causality. Lack of replication does not necessarily imply lack of causality, but might point to the need for more studies in certain populations or more detailed study of the function of a particular gene.

A two-locus model for gene-gene interaction

		Genotype Locus B			Mean
		BB	Bb	bb	
Genotype Locus A	AA	μ_{AABB} $p_A^2 p_B^2$	μ_{AABb} $2p_A^2 p_B p_b$	μ_{AAbb} $p_A^2 p_b^2$	$\mu_{AA\cdot}$
	Aa	μ_{AaBB} $2p_A p_a p_B^2$	μ_{AaBb} $4p_A p_a p_B p_b$	μ_{Aabb} $2p_A p_a p_b^2$	$\mu_{Aa\cdot}$
	aa	μ_{aaBB} $p_a^2 p_B^2$	μ_{aaBb} $2p_a^2 p_B p_b$	μ_{aabb} $p_a^2 p_b^2$	$\mu_{aa\cdot}$
	Mean	$\mu_{\cdot\cdot BB}$	$\mu_{\cdot\cdot Bb}$	$\mu_{\cdot\cdot bb}$	

- Genotypic means and frequencies for a two-locus system

Mean and variances

- The overall mean and variance

$$\mu = p_A^2 p_B^2 \times \mu_{AABB} + 2p_A^2 p_B p_b \times \mu_{AABb} + p_A^2 p_b^2 \times \mu_{AAbb} + 2p_A p_a p_B^2 \times \mu_{AaBB} + 4p_A p_a p_B p_b \times \mu_{AaBb} + 2p_A p_a p_b^2 \times \mu_{Aabb} + p_a^2 p_B^2 \times \mu_{aaBB} + 2p_a^2 p_B p_b \times \mu_{aaBb} + p_a^2 p_b^2 \times \mu_{aabb}.$$

$$\sigma^2 = p_A^2 p_B^2 \times (\mu_{AABB} - \mu)^2 + 2p_A^2 p_B p_b \times (\mu_{AABb} - \mu)^2 + p_A^2 p_b^2 \times (\mu_{AAbb} - \mu)^2 + 2p_A p_a p_B^2 \times (\mu_{AaBB} - \mu)^2 + 4p_A p_a p_B p_b \times (\mu_{AaBb} - \mu)^2 + 2p_A p_a p_b^2 \times (\mu_{Aabb} - \mu)^2 + p_a^2 p_B^2 \times (\mu_{aaBB} - \mu)^2 + 2p_a^2 p_B p_b \times (\mu_{aaBb} - \mu)^2 + p_a^2 p_b^2 \times (\mu_{aabb} - \mu)^2.$$

- The epistatic variance is

$$\sigma_e^2 = \sigma^2 - \sigma_1^2 - \sigma_2^2$$ where

$$\sigma_1^2 = p_A^2 \times (\mu_{AA} - \mu)^2 + 2p_A p_a \times (\mu_{Aa} - \mu)^2 + p_a^2 \times (\mu_{aa} - \mu)^2$$

$$\sigma_2^2 = p_B^2 \times (\mu_{BB} - \mu)^2 + 2p_B p_b \times (\mu_{Bb} - \mu)^2 + p_b^2 \times (\mu_{bb} - \mu)^2$$

Maximum likelihood estimates

- The maximum likelihood estimates (MLEs) can be based on the normal density with overall mean and variance.
- A range of different epistatic models can be compared (Evans et al. 2006).
- The model can be extended to three- or more loci but largely remains of theoretical interest.

Gene-environment interaction (Wikipedia)

- Gene-environment interaction, also called genotype-environment interaction or GxE, is used to describe any phenotypic effects that are due to interactions between the environment and genes.
- Phenylketonuria (PKU) is caused by mutations to a gene coding for a particular liver enzyme. In the absence of this enzyme, an amino acid known as phenylalanine does not get converted into the next amino acid in a biochemical pathway, and therefore too much phenylalanine passes into the blood and other tissues. This disturbs brain development leading to mental retardation and other problems. PKU affects approximately 1 out of every 15,000 infants in the U.S. However, most affected infants do not grow up impaired because of a standard screening program used in the U.S. and other industrialized societies. Newborns found to have high levels of phenylalanine in their blood can be put on a special, phenylalanine-free diet. If they are put on this diet right away and stay on it, these children avoid the severe effects of PKU.

Models for gene-environment interaction

- We employ logistic regression when disease is involved,

$$\Pr(D = 1 \mid G, E) = \frac{e^{\alpha + \gamma_g G + \gamma_e E + \gamma_{ge} GE}}{1 + e^{\alpha + \gamma_g G + \gamma_e E + \gamma_{ge} GE}}$$

- or a linear regression model when quantitative trait is in question,

$$Y = \alpha + \beta_g G + \beta_e E + \beta_{ge} GE + e$$

- There are variations such as unmatched case-control, case-family, case-only designs.

Software

- POWER, QUANTO, power calculation
- ANCESTRYMAP, admixture mapping (also ADMIXMAP but MALDsoft)
- 2LD/GENECOUNTING/HAP, programs for linkage disequilibrium analysis
- HAPLOVIEW, haplotype estimation, visualisation, tagging
- SNPHAP, haplotype estimation involving SNPs
- PHASE, haplotype inference using MCMC and coalescence
- tagSNPs, SNP tagging including generation of SAS program for haplotype trend regression
- FBAT, family-based association tests
- TRANSMIT, haplotype transmission-disequilibrium test (TDT)
- UNPHASED, haplotype analysis involving unrelated individuals or families
- HelixTree, GENOMIZER, PLINK, SNPGWA, tools for GWAS

Retrospective methods

- A paradigm of Cause ← Outcome
- It has link with the retrospective epidemiological design.
- Statistically, this is concerned with P(Y|X) vs P(X|Y), e.g., rarity of disease and use of covariates but there are ample discussions in the literature.
- We show a work by Tan et al. (2007) involving gene-environment interaction. The framework is quite general and invokes the rare disease assumption. It also as link with other work which will be briefly described.

Total accesses to since publication (Oct 2007-Feb2008): 840

Penelope Webb PhD, Biology Editor

Theodora Bloom PhD, Editorial Director

Email: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

Background of retrospective methods

- Prentice RL, Pyke R. Biometrika 1979, **66**:403-11.
- The full retrospective likelihood is proportional to the retrospective likelihood of gene conditional on environmental factors and disease multiplied by the prospective likelihood of disease given genetic factors (Kwee et al. *GE* 2007, **31**:73-90). i.e., $L = P[G,E|D] = P[G|E,D]P[E|D]$
- TDT, Waldman et al. *AHG* 1999, **63**:329-40, Zou GY. *AHG* 2006, **70**:262-72
- Case-control design, e.g. Epstein MP, Satten GA. *AJHG* 2003, **73**:316-29; Weinberg CR, Umbach DM. *AJE* 2000, **152**:197-203, Power and bias were examined by Satten GA, Epstein MP, *GE* 2004, **27**:192-201
- Case-only design, e.g. Khoury & Flanners. *AJE* 1996, **144**:207-13
- Logistic model of discrete and continuous traits for population-based sample
  - Single-locus, Tan et al. *AHG* 2003, **67**:598-607;
  - Haplotype Analysis, Tan et al. *Genet Res* 2005, **86**:223-31
  - GEI, Tan et al. *BMC Genet* 2007, **8**:70
- To relax the independence assumption. Shin J-H, McNeney B, Graham J. SAGMB 2007, **6:**13; Chen et al. *Biostatistics* 2008, **9**:81-99

Statistical models for retrospective analysis

Let

G = genetic variant, 0=non-carriers, 1=carriers

E = environment exposure, 0=non-exposed, 1=exposed

x = trait value

Then, we have

$Logit\ P[G=1,E=1|x]=a_G I+b_E J+(b_G I+b_E J+b_{GxE} IJ)x$

Where *a*=intercept (nuisance parameter), *b*=slope

We can write out $P[G=1,E=1|x]$, $P[G=0,E=0|x]$, $P[G=1,E=0|x]$ and therefore $RR_G(x)=exp(a_G+b_G x)$, $RRR_G(x_2:x_1)=exp[b_G(x_2-x_1)]$

The log-likelihood function is

$$l=\sum_{k=1}^{N}\sum_{i,j=0}^{1} I(G_k=I,E_k=J)p(a_G,a_E,b_G,b_E,b_{G\times E})$$

Some characteristics of retrospective models

- A distinct feature of the model is trait is treated as an independent variable to allow for both categorical and continuous types and no assumption about normality is required (Cordell H. Hum Mol Genet 2002, **11**:2463-8; Hahn et al. Bioinformatics 2003, **19**:376-82).
- Supplementary materials showed that for the case of binary outcome, when the disease is rare, the relative risk estimate is comparable to that from a prospective model.
- It requires interaction variants be independent, otherwise a haplotype analysis model is required (Epstein MP, Satten GA. Am J Hum Genet 2003, **73**:1316-29; Tan et al. Genet Res 2005, **86**:223-31).
- It is equally applicable to study of any main and joint effects models

More discussion (Kwee et al. 2007)

- A prospective likelihood approach by Lake et al. Hum Hered 2003, **55**:56-65 may suffer with respect to retrospective approach (Epstein & Satten 2004). Profile likelihood approaches by Lin et al. Genet Epidemiol 2005, 29:299-312; Spinka et al. Genet Epidemiol 2005, **29**:108-127 require estimating absolute risk of disease from case-control data
- Consider (a) rare disease and (b) haplotype-environment independence. This leads to likelihood-based inference without specifying the distribution of environment covariates in the sample. Following the full likelihood specification, it is shown that $P[E|D]$ contained no information on haplotype and haplotype-environment interaction parameters. Furthermore, case-only study relies on information from $P[G|E,D=1]$
- Simulation studies showed that (with a realistic sample size) under a recessive model both the full retrospective and the prospective approaches yielded flawed results and occasional convergence problem. However, the multiplicative and dominant models give reasonable estimate.
- (To be) implemented in CHAPLIN for case-control haplotype analysis

To wrap up

- Association study currently is the backbone of genetic epidemiology, and therefore a good understanding of the concepts and methods is essential.
- It has higher resolution than linkage and is very powerful.
- The major concerns have been spurious association, and a number of remedies have been proposed.
- It is complementary to GWAS and will remain to play a significant role in gene characterization. The multilocus method (e.g., haplotype analysis), gene-environment interaction and retrospective analysis requires particular attention.
- GWAS and associated issues will be the focus of our next session.

Exercise

- EHPLUS – haplotype estimation and test of case-control association
- ETDT – extended TDT for multiallelic markers
- PLINK – a whole genome association tool set
- HaploView – graphical presentation of LD structure, haplotype-tagging and association testing (+PLINK).
- SAS/Genetics, Stata and R